ПРИКЛАДНАЯ ЛИНГВИСТИКА

Максименко Ольга Ивановна
Olga Maksimenko
Семина Татьяна Алексеевна
Tatiana Semina

Создание корпуса текстов для анализа тональности

Creating a corpus for sentiment analysis

Аннотация: В статье рассмотрены особенности создания корпусов для исследований в области анализа тональности, приведены примеры крупнейших корпусов в этой области. Кроме того, в статье описан корпус статей, собранных автором и показана разметка, которая проводится на данном корпусе. В заключение представлено дальнейшее развитие собираемого корпуса и новые элементы разметки, которые будут добавлены.

Abstract: Peculiarities of creating corpora for investigations in sentiment analysis are considered in this paper and examples of the largest corpora in this field are given. Moreover, an author describes a corpus that was collected and presents the annotation that is being realized in the corpus. In the conclusion the future development of the corpus is declared, as well as new annotation elements that will be added.

Ключевые слова: анализ тональности, корпус, корпусная лингвистика, аннотационная схема, разметка.

Key words: sentiment analysis, corpus, corpus linguistics, annotation schema, tagging.

Анализ тональности, известный как сентимент-анализ или система извлечения мнений – это область изучения мнений, оценок, отношения и эмоций людей по отношению к таким объектам, как продукты, организации, личности, события, проблемы, и их атрибутам. Сентимент (от англ. sentiment – чувство, мнение, настроение) – эмоциональная оценка, выраженная в тексте, также называемая тональностью текста [Пазельская, Соловьев, 2011: 510].

Чаще объектом исследования являются персонализованные тексты: твиты, блоги, отзывы и рецензии, т.е., в первую очередь, тексты, существующие в Интернете [Максименко, 2014]. Интерес к подобному материалу вызван их объемом и содержанием большого количества субъективной информации. Для нашего исследования были выбраны новостные статьи, анализ которых стал активно осуществляться лишь в последнее время. Такие тексты содержат значительное

количество объектов для исследования тональности, в том числе отношения между этими объектами и нередко имплицитно выраженное мнение автора. Эти факторы затрудняют процесс исследования и требуют сбора и разметки отдельных корпусов для создания автоматических систем анализа тональности и для теоретических исследований в этой области.

Тексты, содержащие множество объектов тональности, представляют особую сложность для исследования. На выходе мы хотим получить отношения между сущностями и полярность оценки между ними вида: [A, B, positive / negative / neutral], где A – субъект тональности, B – объект, positive/ negative/ neutral – тональность:

- positive – положительная оценка;
- negative – отрицательная оценка;
- neutral – отсутствие оценки между сущностями [Семина, 2017: 305].

Корпус – представительная совокупность текстов на данном языке [Плунгян, 2008]. Отличием корпусов для анализа тональности является обязательное наличие разметки. В большинстве случаев такие корпусы делают для последующего применения методов моделирования с применением машинного обучения или квантитативного анализа, на неразмеченных данных невозможно в полной мере провести исследование. Корпусы текстов для анализа тональности различаются между собой тематикой текстов, типом текстов и аннотационной схемой.

Для исследования тональности на любом материале необходим корпус текстов, для машинного обучения на прецедентах (например, с алгоритмами Байеса, логистической регрессии или методом опорных векторов) требуется разметка данных. Наш корпус состоит из аналитических статей, поэтому большее внимание мы уделим аналогичным корпусам на других языках, тем не менее, стоит упомянуть ряд известных корпусов с другими текстами, находящихся в свободном доступе.

Тексты Twitter стали в последнее время популярным материалом для анализа тональности из-за широкого разнообразия тем. Одним из наиболее распространенных корпусов с твитами является корпус Sanders Twitter Dataset

[Sanders Analytics Twitter Corpus, URL: https://github.com/zfz/twitter_corpus], распространяющийся через GitHub. Раньше корпус нельзя было скачать в полном объеме напрямую, нужно было скачивать код для установки (на языке Python) и подгружать твиты с разметкой через Twitter API. Сейчас корпус доступен для скачивания в формате cvs без API. Темы твитов данного корпуса: Apple, Google, Twitter, Microsoft. Объем корпуса составляет более 5 тысяч вручную размеченных твитов. В корпусе есть метки positive, negative, neutral и irrelevant. Один из недостатков корпуса, осложняющих анализ, состоит в присутствии твитов не только на английском, но и на испанском и немецком (возможно, присутствуют и другие языки), хотя последние и имеют метку irrelevant. Кроме корпусов твитов об информационных технологиях в свободном доступе находится Health Care Reform Dataset – корпус твитов, собранных во время обсуждения реформы здравоохранения Б. Обамы.

Для русского языка найти корпус твитов сложнее, однако, несколько лет подряд проводятся соревнования SentiRuEval по анализу тональности на материале твитов между исследовательскими группами [SemEval-2016 Task 4: Sentiment Analysis in Twitter, 2016].

Для кинорецензий существует и свободно распространяется корпус Пана [Pang, Lee, 2008], позднее другие исследователи аннотировали часть корпуса с точки зрения тональной релевантности предложений [Scheible, Schutze, 2016].

Для новостных статей была разработана собственная схема разметки и на основе нее собран корпус MPQA. Аннотационная схема доступна для скачивания и встраивается в инструмент для разметки Gate. Последняя версия корпуса (MPQA 3.0) содержит 70 документов. В настоящее время схема аннотации MPQA является одной из наиболее известных, точно так же, как и их корпус размеченных англоязычных статей. Преимущество схемы состоит в возможности связывать субъекты, объекты и текст, содержащий мнение по отношению к каждому из объектов, определении предложения как содержащего сарказм.

Схема MPQA обладает достаточно серьезным недостатком: она не подойдет для статей, используемых в нашей работе, из-за своей чрезмерной сложности. Для коротких

новостных сообщений применить ее в полной мере возможно, для более объемных или аналитических статей, которые не написаны в строго информативном стиле, применение этой схемы представляется невозможным.

Тем не менее, эта схема может оказаться полезной, если сократить количество используемых меток и упростить сам алгоритм разметки. При таком подходе и сохраняется преимущество в виде возможности связывать субъект, объект и текст, содержащий тональность, и разметку возможно применять к собранным статьям. Такой недостаток как сложность схемы и необходимость вручную прописывать все связи между элементами мнения не позволяет решить даже упрощение схемы.

Разметка с помощью аннотационной схемы MPQA получается глубокой, но для аналитических статей она не подходит. Мы пробовали размечать наши статьи в Gate, установив эту схему, однако, даже на коротких статьях и с учетом сокращения оригинальной схемы это не показало хорошего результата. Тем не менее, данная схема вполне применима к другим языкам: для разметки корейского тонального корпуса она была взята за основу, но были добавлены новые атрибуты и изменены старые, что было вызвано особенностями корейского языка. В частности, единицей аннотации у них были не слова, а морфемы.

Для задачи анализа тональности разметка данных имеет большое значение, так как эти данные используются для обучения и тестирования системы. Кроме того, разметка может содержать несколько слоев, то есть, для одного документа можно сделать несколько разметок, каждая из которых может отличаться глубиной разметки и типом размечаемых единиц.

Для наших исследований был собран собственный корпус текстов, на которых проводится разметка и увеличение объема корпуса. Отбираемые в корпус статьи соответствуют ряду критериев:

1) они посвящены аналитическому обзору политики;
2) в одной статье должно быть более 5 неповторяющихся сущностей;
3) одной из сущностей является Россия, возможны эквивалентные номинации.

Статьи различны по объему, но это не является препятствием для исследования и разметки. Статьи взяты с сайта inosmi.ru – сайта с переводами статей разных зарубежных изданий на русский язык.

В настоящее время наш корпус содержит 151 статью, общий объем составляет 158905 слов.

В 102 статьях проведена разметка именованных сущностей. 61 статья имеет разметку именованных сущностей, выполненную автоматически при помощи программы, разработанной факультетом ВМиК МГУ [Алексеев, Лукашевич, 2011], 41 статья содержит ручную разметку именованных сущностей. В инвентарь именованных сущностей входят: {PER}, {GEOPOLIT}, {LOC}, {TITLE}, {EVENT}, {MEDIA} и {ORG} Назначение меток описано в Таблице 1.

Метка	**Назначение**
PER	Человек
GEOPOLIT	Государство как политическое образование
LOC	Местность, локация, город, страна (и т.д.)
TITLE	Название (книги, статьи, доклада и т.д.)
EVENT	Событие
MEDIA	Средства массовой информации
ORG	Организация, структура, членом которой может быть сущность с меткой PER или GEOPOLIT

Таблица 1

Для 84 статей проведена ручная разметка тональных отношений, имеющая следующий вид:

Россия, Дания, neg, current

Россия, НАТО, neg, current

НАТО, Россия, neg, current

В общем виде разметку можно представить следующим образом:

source, target, pos/neg, current/past

- source – автор мнения;

- target – по отношению к кому/чему высказано мнение;
- pos/neg – полярность мнения;
- current/past – актуальность мнения.

Разметка полярности оценки является бинарной, тем не менее, она может применяться и для выявления нейтральной оценки. Обычно под нейтральной оценкой подразумевается, в том числе, отсутствие какой-либо оценки, поэтому возможна простая генерация пар сущностей в каждой статье и проверка их на наличие отношений. Если для пары сущностей не найдено ни положительное, ни отрицательное мнение, можно установить нейтральную оценку между ними.

Актуальность мнения является еще одним важным критерием в анализе тональности, нужно отметить, что этот фактор часто опускают из-за сложности реализации. Мы включили актуальность в разметку, потому что это может быть использовано в будущих исследованиях.

В статьях в нашем корпусе субъектом и объектом мнения почти всегда будут сущности-элементы текста. Для этого может быть важна разметка сущностей, особенно различие между метками GEOPOLIT и LOC. Например, Москва как геополитическая организация может быть источником мнения, в то время как Москва как просто топоним – нет.

29 статей имеют разметку с разрешенной референцией в примерах типа «президент России». При разметке именованных сущностей метка ставится только для России, то есть, речь в высказывании идет о человеке (в данном случае – В.В. Путине), в то время как метка показывает на геополитическое образование. Как говорилось ранее, тип метки может влиять на способность сущности быть источником мнения, поэтому подобная разметка данных может оказаться полезной.

Покрытие корпуса статей разметкой отображено в Таблице 2:

Вид разметки	**Размеченные статьи, %**
Именованные сущности	79,47
Тональные отношения	55,63

Референция в именованных сущностях	19,21

Таблица 2

В будущем планируется продолжение разметки статей, увеличение объема корпуса и добавление новых слоев разметки. Стоит отметить, что полное покрытие статей всеми слоями разметки не является обязательным, для некоторых теоретических исследований и практических разработок будет достаточно и части размеченных данных.

В качестве новых слоев рассматриваются:

- разметка мнений в прямой и косвенной речи;
- выделение сарказма;
- выявление мнения автора.

Работа над последним пунктом ведется в настоящее время, но вместо обычной разметки экспертами мы предлагаем применение одной из метрик информационного поиска [Семина, 2019].

Литература

1. Алексеев, А.А., Лукашевич, Н.В. Автоматическое извлечение сущностей на основе структуры новостного кластера // Искусственный интеллект и принятие решений. 2011. № 4. С. 95–103.
2. Габриелова Е.В., Максименко О.И. Импликация и экспликация оценки в русскоязычном сегменте Твиттера (на примере проблемы миграции)// Ученые записки НОПриЛ 2017 №3 С. 68-75
3. Максименко О.И. Анализ тональности текстов (сентимент-анализ) на материале СМИ // IV Новиковские чтения: Функциональная семантика и семиотика знаковых систем. Сб. научных статей. Часть 1. РУДН, 2014. С. 96–105
4. Пазельская А.Г., Соловьев А.Н. Метод определения эмоций в текстах на русском языке // Компьютерная лингвистика и интеллектуальные технологии: «Диалог-2011». М., 2011. С. 510–522.
5. Плунгян В.А. Корпус как инструмент и как идеология: о некоторых уроках современной корпусной лингвистики //

Русский язык в научном освещении. М., 2008. № 2 (16). С. 7-20.
6. Семина Т.А. Автоматическое выявление отношений между сущностями в мультиобъектных текстах // Человек в информационном пространстве: понимание в коммуникации: сборник научных трудов / под общ.ред. Н.В. Аниськиной, Л.В. Уховой. В 2 тт. Ярославль: Изд-во ЯГПУ, 2017. Т. 1. С. 304 – 309.
7. Семина Т. А. Извлечение мнения автора через обратную частоту документа // Вестник Московского государственного областного университета (электронный журнал). 2019. № 2. URL: www.evestnik-mgou.ru
8. Deng L., Wiebe J. MPQA 3.0: An Entity/Event-Level Sentiment Corpus // Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL. 2015. С. 1323–1328.
9. SemEval-2016 Task 4: Sentiment Analysis in Twitter. 2016 [Электронный ресурс]. – URL:http://alt.qcri.org/semeval2016/task4/data/uploads/semeval2016_task4_report.pdf
10. Pang B., Lee L. Opinion Mining and Sentiment Analysis. In: Foundations and Trends in Information Retrieval, 2008, №. 2. С. 1−135.
11. Sanders Analytics Twitter Corpus [Электронный ресурс] URL: https://github.com/zfz/twitter_corpus (дата обращения: 08.01.2019).
12. Scheible S., Schutze H. Sentiment Relevance // Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia, 2013, С. 954−963.